様式 C-19

科学研究費補助金研究成果報告書

平成　22　年　5　月　7　日現在

研究種目：基盤研究（C）
研究期間：2007～2009
課題番号：19500079
研究課題名（和文）　経歴･オフセット法による多次元データの実装方式とその応用

研究課題名（英文）　An Implementation Scheme of Multidimensional Datasets by History-offset Encoding and Its Applications

研究代表者
都司　達夫（ TATSUO TSUJI ）
福井大学・工学研究科・教授
研究者番号：80115302

研究成果の概要（和文）：
申請者等が提案している経歴･オフセット法による多次元データの実装方式について，つぎの成果を得た。(1)経歴・オフセット空間のオーバフローへの対処法を考究し、それを実装した多次元データ処理基盤を構築し，その有効性を検証した。また，応用として(2)データウェアハウスにおけるカーネルデータ構造としての方式設計とデータキューブの差分構築の方式を提案し、その有効性を検証した。(3)XML 文書木の効率よい，新たなラベル付けの方式を提案し，評価した。

研究成果の概要（英文）：
On the *history-offset* implementation scheme for multidimensional datasets we are proposing, the followings have been achieved: (1) proposed a countermeasure against the overflow of histo-offset space and verified its effectiveness, (2) designed a scheme of kernel data structures employed in data warehouse systems and a scheme of incremental maintenance of datacubes, and (3) proposed and evaluated an efficient labeling scheme for dynamic XML trees.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	1,200,000	360,000	1,560,000
2008 年度	1,200,000	360,000	1,560,000
2009 年度	900,000	270,000	1,170,000
年度			
年度			
総　計	3,300,000	990,000	4,290,000

研究分野：総合領域
科研費の分科・細目：情報学・メディア情報学／データベース
キーワード：　多次元データ，多次元配列，拡張可能配列，経歴・オフセット法

１．研究開始当初の背景

複数の単純型データの組(タップル)からなる多次元データの処理は，近年多次元情報処理の応用分野の拡大につれて，急激にその重要性を増大している。特に，データの大規模化とともにシステム運用時の動的なデータ量増大や構造変化に効率よく対処するための新たな方法論と基盤技術が強く要請されて

いる。

２．研究の目的

我々が最近提案している経歴･オフセット法による動的な多次元データの実装方式は従来方式に比べて時間･空間効率ともに優れた方式である。本研究では，本方式に関する基本研究に基づいて，動的な多次元情報の強化された処理基盤システムを構築すると共に，その上に多次元情報処理の各応用分野の動的特性に基づいたシステムを構築する。従来研究とは異なる視点とアプローチによる取り組みであり，これらの研究を通して得られた，知見に基づき，本方式の新たな応用分野を開拓する。これにより，本方式の広い応用性と有効性を実証する。

本方式は拡張可能配列の考え方に基づいており，データ圧縮と拡張可能性を両立させながら，配列の高速ランダムアクセス機能を阻害しないレコードの encoding，decoding 技法を提案している。関係テーブルのみならず本方式は多次元情報処理における基本実装方式として，きわめて応用性の広い方式であり，この分野において新たな方法論を提案できる可能性を有していると考えられる。

３．研究の方法

関係データベースの実装方式については，現在まで膨大な研究上および実際上の蓄積があるが，申請者等が提案している経歴･オフセット法による関係テーブルの実装方式は従来方式に比べて時間･空間効率ともに極めて優れた方式である。データベース応用のみならず，一般に関係テーブルで表現できるか，わずかの付加的なコストにより，関係テーブルにマッピングできるような多次元アプリケーションデータは広範に存在する。本特定研究期間内においては，データベース応用も含め，これらの応用に対して，本実装方式が有効であることを実証すると共に，新たな応用分野を開拓する。平成 19 年度においては，②で述べた，経歴・オフセット空間のオーバフローの問題を解決すると共に，応用として，OLAP やデータウエアハウス分野への応用を指向して，大規模多次元データのコンパクトな実装方式と高速検索機能および多様な OLAP 処理機能の開発を目標とする。ここではバックエンド DB からの多次元データベースの効率よい構築技法として，申請者等がすでに[2]で提案している，索引部分とデータ部分を分離して構築する技法を利用する。また，平成 20 年度においては，①(1)の応用課題に取り組む。ROLAP(Relational OLAP)におけるデータキューブのインクリメンタルなメンテナンスについてはいくつかの研究があるが，多次元配列を基盤とする MOLAP については申請者が知る限り，ほとんど見受けられない。本研究で実装する経歴・オフセット法による多次元配列の拡張可能性が，非常に効率よく MOLAP におけるデータキューブのインクリメンタルなメンテナンス機能を設計し実装･評価する。さらに，平成 21 年度の前半においては，①(2)の課題に取り組む。経歴オフセット法で実装される拡張可能配列への XML 文書のマッピング手法と更新に強いラベル付けの手法開発およびマッピングされた大規模 XML 木の圧縮格納方式の設計とプロトタイプシステムの実装・評価までを行う。

４．研究成果

［平成19年度の研究成果］

(1) 本方式において最も重要な問題点として，(経歴･オフセット空間のオーバフローへの対処法を考究し、解析評価により、その有効性を検証した[9]。次の2つの対策を提案して，それらの組み合わせにより、対処した。

(a) 多次元データの属性集合を 2 つに分割し，それにしたがって多次元データセットを垂直に 2 つに分割する。これにより，個々の分割が使用している論理空間をそれぞれ独立に確保できる。

(b) 拡張可能配列全体をチャンクと呼ぶ同じ次元数の超立方体に分割し、拡張はこのチャンク単位で行う。これにより、アドレス空間を有効に利用できることになる。

(2) 本方式の重要な応用として，データウェアハウスにおけるカーネルデータ構造としての方式設計とデータキューブの差分構築の問題を扱い、その有効性を解析的に検証した[7][8]。拡張可能配列が動的に追加されるデータ集合を既存の蓄積データを再配置することなく多次元配列に挿入可能であるという、優れた性質を活用して、効率よくデータキューブを差分構築するためのデータ構造とアルゴリズムを、”次元共有“と”部分配列法“の概念の元に設計し評価した。

いずれも，ほぼ当初の計画通りの成果が得られた。

［平成20年度の研究成果］

(1)本年度は経歴･オフセット法に関する前年度における成果に基づいて，データキューブのインクリメタルメンテナンスの方式設計とプロトタイプシステムの実装を行い，評価した[4]。最近の期間内にフロントエンド DB に新たに追加されたレコードのみを対象と

して，既存データキューブとの集約演算を行う。このとき，新たなカラム値を持つレコードに対しては論理拡張可能配列を対応次元方向に拡張する。これにより，事前計算のコストを大幅に削減した。プロトタイプシステムを使用した評価の結果，本提案手法は従来手法に比べて，記憶効率と処理効率の性能が十分高いことを実証した。

(2) 経歴・オフセット法に基づく MOLAP データの差分クラスタリングの方式を提案して，プロトタイプの実装により評価した。

MOLAP システムでは基幹DB に蓄積された関係データベースがダンプされそのファクトテーブルが多次元配列に格納される。多次元配列の高速なランダムアクセス性を生かして，ファクトデータに対して，集約演算をはじめ種々の統計処理を効率よく行うことができる。この高速アクセス性は配列の各次元サイズが固定であることに依っている。ここでは，前回ダンプした配列データを再配置することなしに，差分のみのダンプを行い，多次元配列を構築する。このために，我々が提案している動的な多次元データの実装方式によりシステムを実装する。本論文では，基幹DB とのリアルタイムの一貫性を確保する必要がないことに注目して，クラスタリングされた多次元配列を高速に構築するためのデータバッファリングの方式を提案して評価した[5]。

しかし，この方式では，差分構築前の元の多次元配列データを大量に移動する必要性が高く，差分構築のコストを悪化させていた。ここでは，この悪化を軽減し，差分構築のコストをさらに抑制するための方式を提案して，プロトタイプシステムの構築により評価した。その結果，従来方式よりも低コストで差分クラスタリングが可能であることを示した。さらにこの方式で得られた多次元配列からクラスタリングされた固定サイズの多次元配列を効率よく構築する方式を提案して評価した。

［平成21年度の研究成果］

近年， 半構造データを効率よく扱える言語として，XMLが注目されており，それに伴い， 大規模のXML 文書データを効率よく記憶・管理することの重要性が高まってきている。XML 文書の論理構造は木構造を用いて表現することが可能であるが，XML 文書データの管理において，文書構造を効率よく管理するために，文書から得られるXML 木ノードへの効率のよいラベル付けの方法を提供することが重要である。

経歴・オフセット法の応用として，本年度は，XML 木の構造更新に対応したラベル付けの方式を提案して評価した[1]。この手法は，経歴オフセット法による多次元データのエンコード方式に基づいており，XML 木を多次元拡張可能配列へと埋め込むことによりエンコードを行う。XML木のレベルを拡張可能配列の次元に対応させ，XML木の各ノードを拡張可能配列の当該要素の<経歴，オフセット>対に対応させている。HOMDと呼ぶ，経歴オフセット法の実装データ構造によりXML木を実装しており，XML木のノードの検索は，はHOMDの高速性により，効率よく行える。また，HOMDの拡張性により，XML 木の動的な構造更新に対し，再ラベル付けを行う必要はない。

本方式の最も優れた点として，同種の他方式と比べ，ノードの追加場所と順序に関わらず，ラベルの記憶コストが格段に小さくて済むことが挙げられる。他方式では，同じ場所に複数ノードの追加が起こると，生成されるラベルのサイズは膨大になってしまい，大きな記憶コストが必要となる。提案方式に基づいたプロトタイプシステムを実装し。ラベルサイズ･ラベル記憶コスト･ノード検索コストについて，他方式のDLN， ORDPATH， QED、素数ラベル方式 との比較実験を行い，我々の手法が有効であることを示した。

５．主な発表論文等

〔雑誌論文〕（計 14 件）

① Li B., Kwawaguchi K., Tsuji T., Higuchi K., A Labeling Scheme for Dynamic XML Trees Based on History-offset Encoding, 情報処理学会論文誌 : データベース, Vol. 3, No. 1, pp. 1-17, 2010，査読有

② 都司，水野，松本，樋口，動的多次元データセットのコンパクトな実現方式の提案，日本データベース学科論文誌，Vol.8, No.3, pp. 1-6, 2009，査読有

③ Shimada T., Tsuji T, Higuchi K.， A Secondary Storage Scheme for Multidimensional Data Preserving Proximity, Journal of Digital Information Management, Vol. 7, No.4, , pp.228-236, 2009，査読有

④ Jin D., Tsuji T., Higuch, K.， An Incremental Maintenance Scheme of Data Cubes and Its Evaluation, 情報処理学会論文誌 : データベース, Vol. 1, No. 3, pp. 36-48, 2008，査読有

⑤ 土田, *都司, 樋口, ＭＯＬＡＰ用多次元配列構築のためのバッファリング方式，日本データベース学会論文誌，査読有,Vol.7，No.1, pp.19-24, 2008，査読有

⑥ Tsuchida T., Shimada T., Tsuji T., Higuchi K., Information Storage and Retrieval Schemes for Recycling Products， Journal of Software, , Vol. 3, No. 6, pp. 37-45, 2008.，査読有

⑦]＊ Jin D., Tsuji T., Tsuchida T., Higuch K., An Incremental Maintenance Scheme of Data Cubes, Proc. of Int'l Conference on Database Systems for Advanced Applications (DASFAA

2008), pp.172-187, 2008，査読有
⑧ Tsuji T., Jin D., Higuchi K. , Data Compression for Incremental Data Cube Maintenance, Proc. of Int'l Conference on Database Systems for Advanced Applications (DASFAA 2008) , pp.682-685, 2008，査読有.
⑨Tsuji T., Kuroda M., Higuchi K., History offset implementation scheme for large scale multidimensional data sets, Proc. of the 2008 ACM Symposium on Applied Computing (SAC2008), pp.1021-1028, 2008，査読有
⑩ Li B., Tsuji T., Higuchi K., Sharing Flexibly Resizable Multidimensional Arrays in Client/Server Environment, Proc. of ICDE Workshop，pp. 19-24, 2007，査読有
⑪ Hasan K. M. A., Tsuji T., Higuchi K., An Efficient Implementation for MOLAP Basic Data Structure and Its Evaluation, Proc. of Int'l Conference on Database Systems for Advanced Applications (DASFAA2007), pp.288-299, 2007, 査読有

［注］＊論文[7]は「IPSJ 論文船井若手奨励賞」を受賞した。

〔学会発表〕（計 26 件）
① Jin D., Tsuji T., K.Higuchi K.,, A New Parallel MOLAP Data Cube Construction Scheme 第 2 回データ工学と情報マネジメントに関するフォーラム, F5-4, 2010.
② 嶋田, 佐々原, 都司, 樋口，タプルストリームからのデータキューブの構築第 2 回データ工学と情報マネジメントに関するフォーラム, F5-2, 2010.
③ 小野, 樋口, 都司， XML 文書に対する索引分散管理システムにおける索引分割手法第 1 回データ工学と情報マネジメントに関するフォーラム,
④ Li B., 川口, 都司, 樋口， 構造更新に対応した XML 木のラベル付け方式とその評価データ工学と情報マネジメントに関するフォーラム, C7-4, 2009.
⑤ 土田, 都司, 樋口， MOLAP のための多次元配列のクラスタリングとその評価 データ工学と情報マネジメントに関するフォーラム, E1-6, 2009.
⑥ 嶋田, 都司, 樋口， 次元依存性を考慮した多次元データの二次記憶における圧縮コンテナ化方式 電子情報通信学会 データ工学ワークショップ, D4-1, 2008.
⑦ 川口, Bei Li, 都司, 樋口， 構造更新に対応した XML 木ノードのラベル付けの一方式 電子情報通信学会 データ工学ワークショップ, C8-3, 2008.
⑧ 土田, 都司, 樋口， MOLAP 用多次元配列構築のためのバッファリング方式電子情報通信学会 データ工学ワークショップ, D1-5, 2008.

〔産業財産権〕
○出願状況（計 1 件）
名称：データベース装置，データベースの管理方法，データベースのデータ構造，データベースの管理プログラムおよびそれを記録したコンピュータ読み取り可能な記録媒体，
発明者：都司達夫
権利者：福井大学
種類：特許出願
番号：特願 2009-041176
出願年月日：平成 21 年 2 月 24 日
国内外の別：国内

〔その他〕
ホームページ等
http://www.zakuro.ac.jp

６．研究組織
(1)研究代表者
都司 達夫（TATSUO TSUJI）
福井大学・大学院工学研究科・教授
研究者番号：80115302
(2)研究分担者
樋口 健（KEN HIGUCHI）
福井大学・大学院工学研究科・准教授
研究者番号：50293410